GENERALIZED HIGH-DIMENSIONAL TENSOR LEARNING WITH NUCLEAR NORM REGULARIZATION

Weilin Shen[1], Yao Wang[2,3] and Xuehu Zhu[1]

[1] *School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China*

[2] *School of Management, Xi'an Jiaotong University, Xi'an, China*

[3] *Center for Intelligent Decision-Making and Machine Learning, Xi'an Jiaotong University, Xi'an, China*

Supplementary Material

S1. Conditions for Theorem 2

We outline five conditions to explore the statistical properties that underlie the convergence of our method and to determine the convergence rate of $\widehat{\boldsymbol{\mathcal{M}}}$.

Condition S1.1. $\{(\boldsymbol{\mathcal{X}}_i, \boldsymbol{Y}_i) : i = 1, 2, \dots, n\}$ are i.i.d. from $(\boldsymbol{\mathcal{X}}, \boldsymbol{Y})$, where $vec(\boldsymbol{\mathcal{X}}_i)$ follows a sub-Gaussian distribution such that $\| vec(bcirc(\boldsymbol{\mathcal{X}}_i)) \|_{\psi_2} \leq \kappa_0 < \infty$;

Condition S1.2. The function $b(x)$ satisfies $|b''(x)| \leq M < \infty$ for any

$x \in \mathbb{R}$ and $|b'''(x)| \leq |x|^{-1}$ for $|x| > 1$;

Condition S1.3. The matrix $\boldsymbol{H}(\boldsymbol{\mathcal{M}})$ satisfies $\lambda_{\min}(\boldsymbol{H}(\boldsymbol{\mathcal{M}}^*)) \geq \kappa_l > 0$;

Condition S1.4. $\| \boldsymbol{\mathcal{M}}^* \|_F \geq \alpha\sqrt{dd_3}$ for some constant α;

Condition S1.5. Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ be a 3-D tensor with tSVD decomposition of $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$. If $\boldsymbol{\mathcal{X}} \rightarrow 0$, there exists a constant $c_t > 0$ such that $\| \boldsymbol{\mathcal{S}}(i,:,:) \|_1 \leq c_t \| \boldsymbol{\mathcal{S}}(1,:,:) \|_1, i = 1, 2, \ldots, d_3$.

Remark S1.1. Conditions S1.1–S1.4 are typically necessary for ensuring the consistency of the M-estimate in high-dimensional data, as discussed in Fan et al. [2018]. It's important to note that a majority of real-world datasets adhere to these conditions, and the corresponding $b(\cdot)$ functions in generalized linear regression models, such as the Logistic model, negative binomial regression, Poisson regression, and others, also fulfill the aforementioned bounded conditions.

Condition S1.5 demonstrates the boundedness of the tensor. In reality, when decomposing general real tensor data using tSVD, the resulting f-diagonal tensor possesses elements that are interrelated along the third dimension, and their numerical magnitudes frequently adhere to linear constraints.

S2. Some Useful Lemmas for Tensors with Proofs

Lemma S2.1. *If* $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_1 \times (c+d) \times d_3}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{(c+d) \times d_2 \times d_3}$ *be divided as* $\boldsymbol{\mathcal{A}} = [\boldsymbol{\mathcal{A}}_1, \boldsymbol{\mathcal{A}}_2], \boldsymbol{\mathcal{B}} = \begin{bmatrix} \boldsymbol{\mathcal{B}}_1 \\ \boldsymbol{\mathcal{B}}_2 \end{bmatrix}$ *with* $\boldsymbol{\mathcal{A}}_1 \in \mathbb{R}^{d_1 \times c \times d_3}, \boldsymbol{\mathcal{A}}_2 \in \mathbb{R}^{d_1 \times d \times d_3}, \boldsymbol{\mathcal{B}}_1 \in \mathbb{R}^{c \times d_2 \times d_3}$ *and* $\boldsymbol{\mathcal{B}}_2 \in \mathbb{R}^{d \times d_2 \times d_3}$*, Then we have*

$$\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}} = [\boldsymbol{\mathcal{A}}_1, \boldsymbol{\mathcal{A}}_2] * \begin{bmatrix} \boldsymbol{\mathcal{B}}_1 \\ \boldsymbol{\mathcal{B}}_2 \end{bmatrix} = \boldsymbol{\mathcal{A}}_1 * \boldsymbol{\mathcal{B}}_1 + \boldsymbol{\mathcal{A}}_2 * \boldsymbol{\mathcal{B}}_2.$$

Proof **:**

For $k = 1, 2, \ldots, d_3$, we have

$$\begin{aligned}
(\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}})^{(k)} =& \boldsymbol{A}^{(k)} \boldsymbol{B}^{(1)} + \boldsymbol{A}^{(k-1)} \boldsymbol{B}^{(2)} + \cdots + \boldsymbol{A}^{(k+1)} \boldsymbol{B}^{(d_3)} \\
=& [\boldsymbol{A}_1^{(k)}, \boldsymbol{A}_2^{(k)}] \begin{bmatrix} \boldsymbol{B}_1^{(1)} \\ \boldsymbol{B}_2^{(1)} \end{bmatrix} + \cdots + [\boldsymbol{A}_1^{(k+1)}, \boldsymbol{A}_2^{(k+1)}] \begin{bmatrix} \boldsymbol{B}_1^{(d_3)} \\ \boldsymbol{B}_2^{(d_3)} \end{bmatrix} \\
=& (\boldsymbol{A}_1^{(k)} \boldsymbol{B}_1^{(1)} + \boldsymbol{A}_1^{(k-1)} \boldsymbol{B}_1^{(2)} + \cdots + \boldsymbol{A}_1^{(k+1)} \boldsymbol{B}_1^{(d_3)}) \\
& + (\boldsymbol{A}_2^{(k)} \boldsymbol{B}_2^{(1)} + \boldsymbol{A}_2^{(k-1)} \boldsymbol{B}_2^{(2)} + \cdots + \boldsymbol{A}_2^{(k+1)} \boldsymbol{B}_2^{(d_3)}) \\
=& (\boldsymbol{\mathcal{A}}_1 * \boldsymbol{\mathcal{B}}_1)^{(k)} + (\boldsymbol{\mathcal{A}}_2 * \boldsymbol{\mathcal{B}}_2)^{(k)} \\
=& (\boldsymbol{\mathcal{A}}_1 * \boldsymbol{\mathcal{B}}_1 + \boldsymbol{\mathcal{A}}_2 * \boldsymbol{\mathcal{B}}_2)^{(k)}.
\end{aligned}$$

The proof of this lemma is completed. $\square$

Lemma S2.2. *Let* $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{(a_1+a_2)\times d_2 \times d_3}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{d_2 \times (b_1+b_2)\times d_3}$ *can be divided as* $\boldsymbol{\mathcal{A}} = \begin{bmatrix} \boldsymbol{\mathcal{A}}_1 \\ \boldsymbol{\mathcal{A}}_2 \end{bmatrix}, \boldsymbol{\mathcal{B}} = [\boldsymbol{\mathcal{B}}_1, \boldsymbol{\mathcal{B}}_2]$ *with* $\boldsymbol{\mathcal{A}}_1 \in \mathbb{R}^{a_1 \times d_2 \times d_3}, \boldsymbol{\mathcal{A}}_2 \in \mathbb{R}^{a_2 \times d_2 \times d_3}, \boldsymbol{\mathcal{B}}_1 \in \mathbb{R}^{d_2 \times b_1 \times d_3}$ *and* $\boldsymbol{\mathcal{B}}_2 \in \mathbb{R}^{d_2 \times b_2 \times d_3}$. *Then we have*

$$\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}} = \begin{bmatrix} \boldsymbol{\mathcal{A}}_1 \\ \boldsymbol{\mathcal{A}}_2 \end{bmatrix} * [\boldsymbol{\mathcal{B}}_1, \boldsymbol{\mathcal{B}}_2] = \begin{bmatrix} \boldsymbol{\mathcal{A}}_1 * \boldsymbol{\mathcal{B}}_1 & \boldsymbol{\mathcal{A}}_1 * \boldsymbol{\mathcal{B}}_2 \\ \boldsymbol{\mathcal{A}}_2 * \boldsymbol{\mathcal{B}}_1 & \boldsymbol{\mathcal{A}}_2 * \boldsymbol{\mathcal{B}}_2 \end{bmatrix}.$$

***Proof* :**

For any $k = 1, 2, \ldots, d_3$, we have

$$\begin{aligned}
(\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}})^{(k)} =& \boldsymbol{A}^{(k)}\boldsymbol{B}^{(1)} + \boldsymbol{A}^{(k-1)}\boldsymbol{B}^{(2)} + \cdots + \boldsymbol{A}^{(k+1)}\boldsymbol{B}^{(d_3)} \\
=& \begin{bmatrix} \boldsymbol{A}_1^{(k)} \\ \boldsymbol{A}_2^{(k)} \end{bmatrix} [\boldsymbol{B}_1^{(1)}, \boldsymbol{B}_2^{(1)}] + \cdots + \begin{bmatrix} \boldsymbol{A}_1^{(k+1)} \\ \boldsymbol{A}_2^{(k+1)} \end{bmatrix} [\boldsymbol{B}_1^{(d_3)}, \boldsymbol{B}_2^{(d_3)}] \\
=& \begin{bmatrix} (\boldsymbol{\mathcal{A}}_1 * \boldsymbol{\mathcal{B}}_1)^{(k)} & (\boldsymbol{\mathcal{A}}_1 * \boldsymbol{\mathcal{B}}_2)^{(k)} \\ (\boldsymbol{\mathcal{A}}_2 * \boldsymbol{\mathcal{B}}_1)^{(k)} & (\boldsymbol{\mathcal{A}}_2 * \boldsymbol{\mathcal{B}}_2)^{(k)} \end{bmatrix} \\
=& \begin{bmatrix} \boldsymbol{\mathcal{A}}_1 * \boldsymbol{\mathcal{B}}_1 & \boldsymbol{\mathcal{A}}_1 * \boldsymbol{\mathcal{B}}_2 \\ \boldsymbol{\mathcal{A}}_2 * \boldsymbol{\mathcal{B}}_1 & \boldsymbol{\mathcal{A}}_2 * \boldsymbol{\mathcal{B}}_2 \end{bmatrix}^{(k)}.
\end{aligned}$$

The proof of this lemma is completed. $\square$

Lemma S2.3. $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_1 \times l \times d_3}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{l \times d_2 \times d_3}$, *we have*

$$(\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}})^* = \boldsymbol{\mathcal{B}}^* * \boldsymbol{\mathcal{A}}^*.$$

Proof **:**

$\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}} = (\boldsymbol{\mathcal{B}}^* * \boldsymbol{\mathcal{A}}^*)^*$ equals to $(\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}})^{(k)} = ((\boldsymbol{\mathcal{B}}^* * \boldsymbol{\mathcal{A}}^*)^*)^{(k)}$ for any $k = 1, 2, \dots, d_3$. Because

$$(\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}})^{(k)} = \boldsymbol{A}^{(k)} \boldsymbol{B}^{(1)} + \boldsymbol{A}^{(k-1)} \boldsymbol{B}^{(2)} + \cdots + \boldsymbol{A}^{(k+1)} \boldsymbol{B}^{(d_3)}, \tag{S2.1}$$

and

$$\boldsymbol{\mathcal{B}}^* * \boldsymbol{\mathcal{A}}^* = fold\left(\boldsymbol{C}\right),$$

where

$$\boldsymbol{C} = \begin{bmatrix} (\boldsymbol{B}^{(1)})^* & (\boldsymbol{B}^{(2)})^* & \dots & (\boldsymbol{B}^{(d_3)})^* \\ (\boldsymbol{B}^{(d_3)})^* & (\boldsymbol{B}^{(1)})^* & \dots & (\boldsymbol{B}^{(d_3-1)})^* \\ \vdots & \vdots & \ddots & \vdots \\ (\boldsymbol{B}^{(2)})^* & (\boldsymbol{B}^{(3)})^* & \dots & (\boldsymbol{B}^{(1)})^* \end{bmatrix} * \begin{bmatrix} (\boldsymbol{A}^{(1)})^* \\ (\boldsymbol{A}^{(d_3)})^* \\ \vdots \\ (\boldsymbol{A}^{(2)})^* \end{bmatrix},$$

then it is easy to derive that

$$((\boldsymbol{\mathcal{B}}^* * \boldsymbol{\mathcal{A}}^*)^*)^{(k)} = \boldsymbol{A}^{(k)}\boldsymbol{B}^{(1)} + \boldsymbol{A}^{(k-1)}\boldsymbol{B}^{(2)} + \cdots + \boldsymbol{A}^{(k+1)}\boldsymbol{B}^{(d_3)}, \tag{S2.2}$$

which is equal to (S2.1). This completes the proof of this Lemma. □

Lemma S2.4. *Considering $\boldsymbol{\mathcal{D}}$ and $\boldsymbol{\mathcal{D}}_{\mathcal{W}}$ defined by (2.11) and (2.12) in the main body, we have that $\boldsymbol{\mathcal{D}}_{\mathcal{W}}$ is the projection of $\boldsymbol{\mathcal{D}}$ in $\mathcal{W}$.*

***Proof* :**

To prove this lemma, we need to prove the following two parts: (i). $\boldsymbol{\mathcal{D}}_{\mathcal{W}} \subset \mathcal{W}$; (ii). $\boldsymbol{\mathcal{D}} - \boldsymbol{\mathcal{D}}_{\mathcal{W}} \perp \mathcal{W}$.

Proof of Part (i). According to (2.9) and (2.10) in the main body, we have

$$\begin{aligned}
hor(\boldsymbol{\mathcal{U}}^r) &= \{\boldsymbol{\mathcal{B}}_1 \in \mathbb{R}^{d_1\times 1\times d_3} | \boldsymbol{\mathcal{B}}_1 = \boldsymbol{\mathcal{U}}^r * \boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_1 \in \mathbb{R}^{d_2\times 1\times d_3}\},\\
hor(\boldsymbol{\mathcal{V}}^r) &= \{\boldsymbol{\mathcal{B}}_2 \in \mathbb{R}^{d_1\times 1\times d_3} | \boldsymbol{\mathcal{B}}_2 = \boldsymbol{\mathcal{V}}^r * \boldsymbol{\mathcal{X}}_2, \boldsymbol{\mathcal{X}}_2 \in \mathbb{R}^{d_2\times 1\times d_3}\},\\
lat(\boldsymbol{\mathcal{D}}_{\mathcal{W}}) &= \{\boldsymbol{\mathcal{B}}_3 \in \mathbb{R}^{d_2\times 1\times d_3} | \boldsymbol{\mathcal{B}}_3 = \boldsymbol{\mathcal{D}}_{\mathcal{W}}^* * \boldsymbol{\mathcal{X}}_3, \boldsymbol{\mathcal{X}}_3 \in \mathbb{R}^{d_1\times 1\times d_3}\},\\
hor(\boldsymbol{\mathcal{D}}_{\mathcal{W}}) &= \{\boldsymbol{\mathcal{B}}_4 \in \mathbb{R}^{d_1\times 1\times d_3} | \boldsymbol{\mathcal{B}}_4 = \boldsymbol{\mathcal{D}}_{\mathcal{W}} * \boldsymbol{\mathcal{X}}_4, \boldsymbol{\mathcal{X}}_4 \in \mathbb{R}^{d_2\times 1\times d_3}\}.
\end{aligned}$$

For any $\boldsymbol{\mathcal{A}}_1 \in lat(\boldsymbol{\mathcal{D}}_{\mathcal{W}}), \boldsymbol{\mathcal{A}}_2 \in hor(\boldsymbol{\mathcal{D}}_{\mathcal{W}})$, we have

$$\begin{aligned}
\boldsymbol{\mathcal{A}}_1 &= \boldsymbol{\mathcal{D}}_{\mathcal{W}}^* * \boldsymbol{\mathcal{Y}}_1 = (\boldsymbol{\mathcal{U}}^r * \boldsymbol{\mathcal{S}}_{11} * (\boldsymbol{\mathcal{V}}^r)^*)^* * \boldsymbol{\mathcal{Y}}_1 \\
&= \boldsymbol{\mathcal{V}}^r * (\boldsymbol{\mathcal{S}}_{11}^* * (\boldsymbol{\mathcal{U}}^r)^* * \boldsymbol{\mathcal{Y}}_1) \in hor(\boldsymbol{\mathcal{V}}^r), \\
\boldsymbol{\mathcal{A}}_2 &= \boldsymbol{\mathcal{D}}_{\mathcal{W}} * \boldsymbol{\mathcal{Y}}_2 = \boldsymbol{\mathcal{U}}^r * \boldsymbol{\mathcal{S}}_{11} * (\boldsymbol{\mathcal{V}}^r)^* * \boldsymbol{\mathcal{Y}}_2 \\
&= \boldsymbol{\mathcal{U}}^r * (\boldsymbol{\mathcal{S}}_{11} * (\boldsymbol{\mathcal{V}}^r)^* * \boldsymbol{\mathcal{Y}}_2) \in hor(\boldsymbol{\mathcal{U}}^r),
\end{aligned}$$

where $\boldsymbol{\mathcal{Y}}_1 \in \mathbb{R}^{d_1 \times 1 \times d_3}, \boldsymbol{\mathcal{Y}}_2 \in \mathbb{R}^{d_2 \times 1 \times d_3}$. Thus, $\boldsymbol{\mathcal{D}}_{\mathcal{W}} \subset \mathcal{W}$ is proved. We have proved the first part of the lemma.

Proof of Part (ii). For any $\boldsymbol{\mathcal{A}} \in \mathcal{W}$, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times 1 \times d_3}$ and $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{d_2 \times 1 \times d_3}$, there exist $\boldsymbol{\mathcal{X}}_0, \boldsymbol{\mathcal{Y}}_0 \in \mathbb{R}^{d_2 \times 1 \times d_3}$ satisfying

$$\boldsymbol{\mathcal{A}}^* * \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{V}}^r * \boldsymbol{\mathcal{X}}_0, \tag{S2.3}$$

$$\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{U}}^r * \boldsymbol{\mathcal{Y}}_0. \tag{S2.4}$$

Altogether the results of Lemma S2.2 and Lemma S2.3, it can be easily proved that there exists $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{r \times r \times d_3}$ such that $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}}^r * \boldsymbol{\mathcal{Z}} * (\boldsymbol{\mathcal{V}}^r)^*$.

Therefore, we have

$$\begin{aligned}\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{D}} - \boldsymbol{\mathcal{D}}_{\mathcal{W}} \rangle =& \langle \boldsymbol{\mathcal{U}}^r * \boldsymbol{\mathcal{Z}} * (\boldsymbol{\mathcal{V}}^r)^*, \boldsymbol{\mathcal{U}}^{r\perp} * \boldsymbol{\mathcal{S}}_{21} * (\boldsymbol{\mathcal{V}}^r)^* + \boldsymbol{\mathcal{U}}^r * \boldsymbol{\mathcal{S}}_{12} * (\boldsymbol{\mathcal{V}}^{r\perp})^* \\ &+ \boldsymbol{\mathcal{U}}^{r\perp} * \boldsymbol{\mathcal{S}}_{22} * (\boldsymbol{\mathcal{V}}^{r\perp})^* \rangle \\ =& 0. \end{aligned} \tag{S2.5}$$

Because $\boldsymbol{\mathcal{A}}$ is arbitrary, we have $\boldsymbol{\mathcal{D}} - \boldsymbol{\mathcal{D}}_{\mathcal{W}} \perp \mathcal{W}$. This completes the proof.

□

Lemma S2.5. *For any tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d \times d \times d_3}$, the tubal rank of $\boldsymbol{\mathcal{A}}_{\overline{\mathcal{W}}}$ is no more than $2r$, where $\mathcal{W}$ and r are defined by (2.7) in the main body.*

***Proof* :**

As $\boldsymbol{\mathcal{A}}_{\overline{\mathcal{W}}}$ can be writen as follow:

$$\boldsymbol{\mathcal{A}}_{\overline{\mathcal{W}}} = [\boldsymbol{\mathcal{U}}^r, \boldsymbol{\mathcal{U}}^{r\perp}] * \begin{bmatrix} \boldsymbol{\mathcal{S}}_{11} & \boldsymbol{\mathcal{S}}_{12} \\ \boldsymbol{\mathcal{S}}_{21} & 0 \end{bmatrix} * [\boldsymbol{\mathcal{V}}^r, \boldsymbol{\mathcal{V}}^{r\perp}]^*,$$

we have

$$rank(\boldsymbol{\mathcal{A}}_{\overline{\mathcal{W}}}) = rank\left(\begin{bmatrix} \boldsymbol{\mathcal{S}}_{11} & \boldsymbol{\mathcal{S}}_{12} \\ \boldsymbol{\mathcal{S}}_{21} & 0 \end{bmatrix}\right).$$

Denote that $\boldsymbol{\mathcal{S}} = \begin{bmatrix} \boldsymbol{\mathcal{S}}_{11} & \boldsymbol{\mathcal{S}}_{12} \\ \boldsymbol{\mathcal{S}}_{21} & 0 \end{bmatrix}$, $\boldsymbol{\mathcal{S}}^{(i)} = \begin{bmatrix} \boldsymbol{\mathcal{S}}_{11}^{(i)} & \boldsymbol{\mathcal{S}}_{12}^{(i)} \\ \boldsymbol{\mathcal{S}}_{21}^{(i)} & 0 \end{bmatrix}$, $i = 1, 2, \dots, d_3$. Then we have $rank(\boldsymbol{\mathcal{S}}^{(i)}) \leq 2r$. Because Discrete Fourier Transformation is invertible, we have $rank(\overline{\boldsymbol{\mathcal{S}}}^{(i)}) = rank(\boldsymbol{\mathcal{S}}^{(i)}) \leq 2r$. Therefore,

$$rank(\boldsymbol{\mathcal{A}}_{\overline{\mathcal{W}}}) = rank(\boldsymbol{\mathcal{S}}) = \max_{i=1,\dots,d_3} rank(\overline{\boldsymbol{\mathcal{S}}}^{(i)}) \leq 2r.$$

This completes the proof. $\square$

Lemma S2.6. *For any tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, we have*

$$\| \boldsymbol{\mathcal{A}} \|_F \leq \sqrt{d_3} \| \boldsymbol{\mathcal{A}} \|_* .$$

***Proof* :**

Considering the provided tensor $\boldsymbol{\mathcal{A}}$, we have

$$\| \boldsymbol{\mathcal{A}} \|_F = \frac{1}{\sqrt{d_3}} \| \overline{\boldsymbol{A}} \|_F \leq \frac{1}{\sqrt{d_3}} \| \overline{\boldsymbol{A}} \|_* = \sqrt{d_3} \| \boldsymbol{\mathcal{A}} \|_* .$$

This completes the proof. $\square$

Lemma S2.7. *For any tensors $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, if $\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle = 0$, we have*

$$\| \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}} \|_* = \| \boldsymbol{\mathcal{A}} \|_* + \| \boldsymbol{\mathcal{B}} \|_* .$$

***Proof* :**

For any $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$, we have

$$\begin{aligned}
\| \boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}} \|_* &= \frac{1}{d_3} \| bcirc(\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}) \|_* \\
&= \frac{1}{d_3} \| bcirc(\boldsymbol{\mathcal{A}}) + bcirc(\boldsymbol{\mathcal{B}}) \|_* \\
&= \frac{1}{d_3} \| bcirc(\boldsymbol{\mathcal{A}}) \|_* + \frac{1}{d_3} \| bcirc(\boldsymbol{\mathcal{B}}) \|_* \\
&= \| \boldsymbol{\mathcal{A}} \|_* + \| \boldsymbol{\mathcal{B}} \|_* .
\end{aligned}$$

The proof is finished. □

Lemma S2.8. *For any tensors $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, we have $\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle \leq \| \boldsymbol{\mathcal{A}} \|_{op} \| \boldsymbol{\mathcal{B}} \|_*$.*

***Proof* :**

For any $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$, we have

$$\begin{aligned}
\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle &= \frac{1}{d_3} \langle bcirc(\boldsymbol{\mathcal{A}}), bcirc(\boldsymbol{\mathcal{B}}) \rangle \\
&\leq \| bcirc(\boldsymbol{\mathcal{A}}) \|_{op} \left(\frac{1}{d_3} \| bcirc(\boldsymbol{\mathcal{B}}) \|_*\right) \\
&= \| \boldsymbol{\mathcal{A}} \|_{op} \| \boldsymbol{\mathcal{B}} \|_* .
\end{aligned}$$

The proof is concluded. □

S3. Proof of Theorem 2

In this part, we propose the following two lemmas to help justify Theorem 2. The proof details of the two lemmas are given in the supplementary material.

Lemma S3.1. *Under the Conditions S1.1 and S1.2, for any* $\nu > 0$*, there exists some constant* $\gamma > 0$ *satisfying if* $\frac{dd_3}{n} < \gamma$*, the following inequality holds for some constants* $C^{(1)}$ *and* $c^{(1)}$*:*

$$\mathbb{P}\left(\| \frac{1}{n} \sum_{i=1}^{n} [b'(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle) - \boldsymbol{Y}_i] \cdot \boldsymbol{\mathcal{X}}_i \|_{op} > \nu \sqrt{\frac{dd_3}{n}} \right) \leq C^{(1)} \exp(-c^{(1)} dd_3).$$

Lemma S3.2. *Under the Conditions S1.1–S1.5, assume* $\lambda \geq \nu \sqrt{\frac{dd_3}{n}}$ *with* ν *defined in Lemma S3.1. Define* $\mathcal{N} := \{\boldsymbol{\mathcal{M}} \in \mathbb{R}^{d \times d \times d_3} :\| \boldsymbol{\mathcal{M}} - \boldsymbol{\mathcal{M}}^* \|_F^2 \leq C_1^{(2)} \rho \lambda^{2-q}, \boldsymbol{\mathcal{M}} - \boldsymbol{\mathcal{M}}^* \in \mathcal{C}(\mathcal{W}, \overline{\mathcal{W}}^{\perp}, \boldsymbol{\mathcal{M}}^*)\}$*. If* $\rho \lambda^{1-q}$ *is small enough,* $L(\boldsymbol{\mathcal{M}})$ *is such that* $LRSC(\mathcal{C}(\mathcal{W}, \overline{\mathcal{W}}^{\perp}, \boldsymbol{\mathcal{M}}^*), \mathcal{N}, \kappa, \tau_\ell)$ *with a probability at least* $1 - C_2^{(2)} exp(-c_1^{(2)} dd_3)$*, where* $\tau_\ell = C_0^{(2)} \rho \lambda^{2-q}, 0 < \kappa < \kappa_\ell$ *and* $c_1^{(2)}, C_0^{(2)}, C_1^{(2)}$ *and* $C_2^{(2)}$ *are constants.*

According to Lemma S3.1 and $\lambda = 2\nu \sqrt{dd_3/n}$, we can learn that there exist constants C_4 and c_1 satisfying

$$\begin{aligned}
&\mathbb{P}\Bigg(\lambda < 2 \parallel \frac{1}{n}\sum_{i=1}^{n}[b'(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle) - \boldsymbol{Y}_i] \cdot \boldsymbol{\mathcal{X}}_i \parallel_{op} \Bigg) \\
=&\mathbb{P}\Bigg(2\nu\sqrt{dd_3/n} < 2 \parallel \frac{1}{n}\sum_{i=1}^{n}[b'(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle) - \boldsymbol{Y}_i] \cdot \boldsymbol{\mathcal{X}}_i \parallel_{op} \Bigg) \\
=&\mathbb{P}\Bigg(\parallel \frac{1}{n}\sum_{i=1}^{n}[b'(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle) - \boldsymbol{Y}_i] \cdot \boldsymbol{\mathcal{X}}_i \parallel_{op} > \nu\sqrt{\frac{dd_3}{n}}\Bigg) \\
\leq& C_4 \exp(-c_1 dd_3). \qquad \text{(S3.6)}
\end{aligned}$$

Then by Lemma S3.2, there are some constants C_1, c_2 and C_5, while $\rho d_3(dd_3/n)^{(1-q)/2} \leq C_1$, we can get

$$\begin{aligned}
&\mathbb{P}\Bigg(L(\boldsymbol{\mathcal{M}}) \text{ doesn't satisfy } LRSC(\mathcal{C}(\mathcal{W}, \overline{\mathcal{W}}^{\perp}, \boldsymbol{\mathcal{M}}^*), \mathcal{N}, \kappa, \tau_\ell)\Bigg) \\
&\qquad\qquad \leq C_5 exp(-c_2 dd_3). \qquad \text{(S3.7)}
\end{aligned}$$

Combining (S3.6) and (S3.7), we can obtain that

$$
\begin{aligned}
&\mathbb{P}\Big(\lambda \geq 2 \parallel \frac{1}{n}\sum_{i=1}^{n}[b'(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle) - \boldsymbol{Y}_i] \cdot \boldsymbol{\mathcal{X}}_i \parallel_{op} \\
&\quad \text{and } L(\boldsymbol{\mathcal{M}}) \text{ satisfy } LRSC(\mathcal{C}(\mathcal{W}, \overline{\mathcal{W}}^{\perp}, \boldsymbol{\mathcal{M}}^*), \mathcal{N}, \kappa, \tau_\ell)\Big) \\
=&1 - \mathbb{P}\Big(\lambda < 2 \parallel \frac{1}{n}\sum_{i=1}^{n}[b'(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle) - \boldsymbol{Y}_i] \cdot \boldsymbol{\mathcal{X}}_i \parallel_{op} \\
&\quad \text{or } L(\boldsymbol{\mathcal{M}}) \text{ doesn't satisfy } LRSC(\mathcal{C}(\mathcal{W}, \overline{\mathcal{W}}^{\perp}, \boldsymbol{\mathcal{M}}^*), \mathcal{N}, \kappa, \tau_\ell)\Big) \\
\geq&1 - \mathbb{P}\Big(\lambda < 2 \parallel \frac{1}{n}\sum_{i=1}^{n}[b'(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle) - \boldsymbol{Y}_i] \cdot \boldsymbol{\mathcal{X}}_i \parallel_{op}\Big) \\
&\quad - \mathbb{P}\Big(L(\boldsymbol{\mathcal{M}}) \text{ doesn't satisfy } LRSC(\mathcal{C}(\mathcal{W}, \overline{\mathcal{W}}^{\perp}, \boldsymbol{\mathcal{M}}^*), \mathcal{N}, \kappa, \tau_\ell)\Big) \\
\geq&1 - C_4 \exp(-c_1 d d_3) - C_5 exp(-c_2 d d_3).
\end{aligned} \tag{S3.8}
$$

(S3.8) reveals that $\lambda \geq 2 \parallel \frac{1}{n}\sum_{i=1}^{n}[b'(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle) - \boldsymbol{Y}_i] \cdot \boldsymbol{\mathcal{X}}_i \parallel_{op}$ and $L(\boldsymbol{\mathcal{M}})$ satisfy $LRSC(\mathcal{C}(\mathcal{W}, \overline{\mathcal{W}}^{\perp}, \boldsymbol{\mathcal{M}}^*), \mathcal{N}, \kappa, \tau_\ell)$ are established at the same time with a high probability of at least $1 - C_4 \exp(-c_1 d d_3) - C_5 exp(-c_2 d d_3)$. So that, it remind us that if we can prove the convergence of our method in the above situations, then we can consider our method to be able to converge with high probability. In the following proof, we shall demonstrate this fact in detail.

Let $\widehat{\boldsymbol{\mathcal{M}}}_t = \boldsymbol{\mathcal{M}}^* + t(\widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}^*)$, where $t = 1$ if $\parallel \widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}^* \parallel_F \leq \ell$, and

$t = \ell / \parallel \widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}^* \parallel_F$ if $\parallel \widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}^* \parallel_F > \ell$. Let $\widehat{\boldsymbol{\mathcal{D}}} = \widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}^*, \widehat{\boldsymbol{\mathcal{D}}}_t = \widehat{\boldsymbol{\mathcal{M}}}_t - \boldsymbol{\mathcal{M}}^*$.

Applying Lemma S2.7, we have

$$\begin{aligned}
\parallel \boldsymbol{\mathcal{M}}^* + \boldsymbol{\mathcal{D}} \parallel_* &\geq \parallel \boldsymbol{\mathcal{M}}^*_{\mathcal{W}} + \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}^\perp} \parallel_* - \parallel \boldsymbol{\mathcal{M}}^*_{\mathcal{W}^\perp} + \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}} \parallel_* \\
&\geq \parallel \boldsymbol{\mathcal{M}}^*_{\mathcal{W}} + \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}^\perp} \parallel_* - \parallel \boldsymbol{\mathcal{M}}^*_{\mathcal{W}^\perp} \parallel_* - \parallel \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}} \parallel_* \\
&= \parallel \boldsymbol{\mathcal{M}}^*_{\mathcal{W}} \parallel_* + \parallel \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}^\perp} \parallel_* - \parallel \boldsymbol{\mathcal{M}}^*_{\mathcal{W}^\perp} \parallel_* - \parallel \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}} \parallel_* .
\end{aligned}$$

As $\parallel \boldsymbol{\mathcal{M}}^* \parallel_* \leq \parallel \boldsymbol{\mathcal{M}}^*_{\mathcal{W}} \parallel_* + \parallel \boldsymbol{\mathcal{M}}^*_{\mathcal{W}^\perp} \parallel_*$, we have

$$\parallel \boldsymbol{\mathcal{M}}^* + \boldsymbol{\mathcal{D}} \parallel_* - \parallel \boldsymbol{\mathcal{M}}^* \parallel_* \geq \parallel \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}^\perp} \parallel_* - \parallel \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}} \parallel_* - 2 \parallel \boldsymbol{\mathcal{M}}^*_{\mathcal{W}^\perp} \parallel_* . \quad \text{(S3.9)}$$

Due to the results of Lemma S2.8 and the convexity of $L(\cdot)$, we get that

$$\begin{aligned}
L(\boldsymbol{\mathcal{M}}^* + \boldsymbol{\mathcal{D}}) - L(\boldsymbol{\mathcal{M}}^*) &\geq -|\langle \nabla L(\boldsymbol{\mathcal{M}}^*), \boldsymbol{\mathcal{D}} \rangle| \\
&\geq - \parallel \nabla L(\boldsymbol{\mathcal{M}}^*) \parallel_{op} \parallel \boldsymbol{\mathcal{D}} \parallel_* \\
&\geq -\frac{\lambda}{2} \parallel \boldsymbol{\mathcal{D}} \parallel_* \\
&\geq -\frac{\lambda}{2} (\parallel \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}^\perp} \parallel_* + \parallel \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}} \parallel_*). \qquad \text{(S3.10)}
\end{aligned}$$

We calculate λ(S3.9)+(S3.10) and have

$$L(\boldsymbol{\mathcal{M}}^* + \boldsymbol{\mathcal{D}}) + \lambda \parallel \boldsymbol{\mathcal{M}}^* + \boldsymbol{\mathcal{D}} \parallel_* - (L(\boldsymbol{\mathcal{M}}^*) + \lambda \parallel \boldsymbol{\mathcal{M}}^* \parallel_*)$$
$$\geq \frac{\lambda}{2}(\parallel \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}^\perp} \parallel_* - 3 \parallel \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}} \parallel_* - 4 \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*)). \tag{S3.11}$$

Obviously, replacing $\boldsymbol{\mathcal{D}}$ with $\widehat{\boldsymbol{\mathcal{D}}}$, the left part of (S3.11) ≤ 0. Then we refer that if $\lambda \geq 2 \parallel \frac{1}{n}\sum_{i=1}^n [b'(\langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{M}}^* \rangle) - \boldsymbol{Y}_i] \cdot \boldsymbol{\mathcal{X}}_i \parallel_{op}$, $\widehat{\boldsymbol{\mathcal{D}}}$ belongs to the following set:

$$\mathcal{C}(\mathcal{W}, \overline{\mathcal{W}}^\perp, \boldsymbol{\mathcal{M}}^*) := \{\parallel \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}^\perp} \parallel_* \leq 3 \parallel \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}} \parallel_* + 4 \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*)\}. \tag{S3.12}$$

Additionally, $\widehat{\boldsymbol{\mathcal{D}}}_t$ alse falls in this set because $\widehat{\boldsymbol{\mathcal{D}}}_t$ is parallel to $\widehat{\boldsymbol{\mathcal{D}}}$. Since $\parallel \widehat{\boldsymbol{\mathcal{D}}}_t \parallel_* \leq \ell$ and $L(\boldsymbol{\mathcal{M}})$ satisfies $LRSC(\mathcal{C}, \mathcal{N}, \kappa_\ell, \tau_\ell)$, we can prove that

$$\kappa_\ell \parallel \widehat{\boldsymbol{\mathcal{D}}}_t \parallel_F^2 - \tau_\ell \leq \langle \nabla L(\widehat{\boldsymbol{\mathcal{M}}}_t) - \nabla L(\boldsymbol{\mathcal{M}}^*), \widehat{\boldsymbol{\mathcal{D}}}_t \rangle =: D_L^s(\widehat{\boldsymbol{\mathcal{M}}}_t, \boldsymbol{\mathcal{M}}^*), \tag{S3.13}$$

where $D_L(\boldsymbol{\mathcal{M}}_1, \boldsymbol{\mathcal{M}}_2) = L(\boldsymbol{\mathcal{M}}_1) - L(\boldsymbol{\mathcal{M}}_2) - \langle \nabla L(\boldsymbol{\mathcal{M}}_2, \boldsymbol{\mathcal{M}}_1 - \boldsymbol{\mathcal{M}}_2) \rangle \geq 0$ and $D_L^s(\boldsymbol{\mathcal{M}}_1, \boldsymbol{\mathcal{M}}_2) = D_L(\boldsymbol{\mathcal{M}}_1, \boldsymbol{\mathcal{M}}_2) + D_L(\boldsymbol{\mathcal{M}}_2, \boldsymbol{\mathcal{M}}_1)$, with help of the convexity of the function $L(\cdot)$.

Define $Q(t) = D_L(\widehat{\boldsymbol{\mathcal{M}}}_t, \boldsymbol{\mathcal{M}}^*)$. It is easy to prove that $Q(t)$ is convex and $Q'(t) = \langle \nabla L(\widehat{\boldsymbol{\mathcal{M}}}_t) - \nabla L(\boldsymbol{\mathcal{M}}^*), \widehat{\boldsymbol{\mathcal{D}}} \rangle = \frac{1}{t} D_L^s(\widehat{\boldsymbol{\mathcal{M}}}_t, \boldsymbol{\mathcal{M}}^*), 0 < t \leq 1$, which yield that

$$D_L^s(\widehat{\boldsymbol{\mathcal{M}}}_t, \boldsymbol{\mathcal{M}}^*) = tQ'(t) \leq tQ'(1) = tD_L^s(\widehat{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{M}}^*).$$

Then, using the above inequalities in (S3.13), we conclude that

$$\begin{aligned} \kappa_\ell \parallel \widehat{\boldsymbol{\mathcal{D}}}_t \parallel_F^2 - \tau_\ell \leq & D_L^s(\widehat{\boldsymbol{\mathcal{M}}}_t, \boldsymbol{\mathcal{M}}^*) \\ \leq & tD_L^s(\widehat{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{M}}^*) \\ = & \langle \nabla L(\widehat{\boldsymbol{\mathcal{M}}}) - \nabla L(\boldsymbol{\mathcal{M}}^*), \widehat{\boldsymbol{\mathcal{D}}}_t \rangle. \end{aligned} \tag{S3.14}$$

For $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d \times d \times d_3}$, when $rank(\boldsymbol{\mathcal{A}}) \leq r$, we have

$$\parallel \boldsymbol{\mathcal{A}} \parallel_* = \frac{1}{d_3} \parallel \overline{\boldsymbol{A}} \parallel_* \leq \frac{\sqrt{rd_3}}{d_3} \parallel \overline{\boldsymbol{A}} \parallel_F = \sqrt{r} \parallel \boldsymbol{\mathcal{A}} \parallel_F .$$

Due to the fact that $\widehat{\boldsymbol{\mathcal{M}}}$ is the solution of optimization problem (3.16) in main body, we can easily refer that $\nabla L(\widehat{\boldsymbol{\mathcal{M}}}) + \lambda \boldsymbol{\mathcal{E}} = 0$, where $\boldsymbol{\mathcal{E}}$ is the subgradient of the $\parallel \boldsymbol{\mathcal{M}} \parallel_*$ at $\boldsymbol{\mathcal{M}} = \widehat{\boldsymbol{\mathcal{M}}}$. It is known from Theorem 3.2 of Lu et al. [2020] that $\boldsymbol{\mathcal{E}} = \{\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^* + \boldsymbol{\mathcal{W}} | \boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{W}} = 0, \boldsymbol{\mathcal{W}} * \boldsymbol{\mathcal{V}} = 0, \parallel \boldsymbol{\mathcal{W}} \parallel_{op} = 1\}$, where $\widehat{\boldsymbol{\mathcal{M}}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$. Altogether the result in (S3.14), we have

$$\begin{aligned}
\kappa_\ell \parallel \widehat{\boldsymbol{\mathcal{D}}}_t \parallel_F^2 - \tau_\ell &\leq -\langle \nabla L(\widehat{\boldsymbol{\mathcal{M}}}^*) + \lambda \boldsymbol{\mathcal{E}}, \widehat{\boldsymbol{\mathcal{D}}}_t \rangle \\
&\leq \parallel \nabla L(\widehat{\boldsymbol{\mathcal{M}}}^*) + \lambda \boldsymbol{\mathcal{E}} \parallel_{op} \parallel \widehat{\boldsymbol{\mathcal{D}}}_t \parallel_* \\
&\leq (\parallel \nabla L(\widehat{\boldsymbol{\mathcal{M}}}^*) \parallel_{op} + \lambda \parallel_{op} \boldsymbol{\mathcal{E}} \parallel_{op}) \parallel \widehat{\boldsymbol{\mathcal{D}}}_t \parallel_* \\
&\leq [\frac{\lambda}{2} + \lambda(\parallel \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{V}}^* \parallel_{op} + \parallel \boldsymbol{\mathcal{W}} \parallel_{op})] \parallel \widehat{\boldsymbol{\mathcal{D}}}_t \parallel_* \\
&\leq 2.5\lambda \parallel \widehat{\boldsymbol{\mathcal{D}}}_t \parallel_* \\
&\leq 10\lambda \parallel (\widehat{\boldsymbol{\mathcal{D}}}_t)_{\overline{\mathcal{W}}} \parallel_* + 10\lambda \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*) \\
&\leq 10\lambda\sqrt{2r} \parallel (\widehat{\boldsymbol{\mathcal{D}}}_t)_{\overline{\mathcal{W}}} \parallel_F + 10\lambda \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*) \\
&\leq 10\lambda\sqrt{2r} \parallel \widehat{\boldsymbol{\mathcal{D}}}_t \parallel_F + 10\lambda \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*). \qquad \text{(S3.15)}
\end{aligned}$$

Choosing $r = \#\{j \in \{1, 2, \ldots, d\} | \sigma_j(\boldsymbol{\mathcal{M}}^*) \geq \tau\}$ with $\tau > 0$, it can be easily promoted that

$$\sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*) \leq \tau^{1-q} \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*)^q \leq \tau^{1-q}\rho. \qquad \text{(S3.16)}$$

At the same time, $\rho \geq \sum_{j \leq r} \sigma_j(\boldsymbol{\mathcal{M}}^*)^q \geq r\tau^q$, then we have $r \leq \rho\tau^{-q}$. Setting $\tau = \lambda/\kappa_\ell$, the results of (S3.15) and (S3.16) and $\tau_\ell = C_0\rho\lambda^{2-q}/\kappa_\ell^{1-q}$ reveal that for some constant C_1^*, $\parallel \widehat{\boldsymbol{\mathcal{D}}}_t \parallel_F \leq C_1^* \sqrt{\rho}(\frac{\lambda}{\kappa_\ell})^{1-\frac{q}{2}}$. If we choose

$l > C_1^* \sqrt{\rho} (\frac{\lambda}{\kappa_\ell})^{1-\frac{q}{2}}$ in advance, we have $\boldsymbol{\mathcal{D}}_t = \boldsymbol{\mathcal{D}}$. Via an application of Lemma S2.5, we can easily get $rank(\widehat{\boldsymbol{\mathcal{D}}}_{\overline{\mathcal{W}}}) \leq 2r$. Consequently, we have

$$
\begin{aligned}
\| \widehat{\boldsymbol{\mathcal{D}}} \|_* &\leq \| \widehat{\boldsymbol{\mathcal{D}}}_{\overline{\mathcal{W}}} \|_* + \| \widehat{\boldsymbol{\mathcal{D}}}_{\overline{\mathcal{W}}^\perp} \|_* \\
&\leq 4 \| \widehat{\boldsymbol{\mathcal{D}}}_{\overline{\mathcal{W}}} \|_* + 4 \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*) \\
&\leq 4\sqrt{2r} \| \widehat{\boldsymbol{\mathcal{D}}}_{\overline{\mathcal{W}}} \|_F + 4 \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*) \\
&\leq 4\sqrt{2r} \| \widehat{\boldsymbol{\mathcal{D}}} \|_F + 4 \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*) \\
&\leq 4\sqrt{2\rho} \tau^{1-\frac{q}{2}} C_1^* \sqrt{\rho} (\frac{\lambda}{\kappa_\ell})^{1-\frac{q}{2}} + 4\tau^{1-q}\rho \\
&= (4\sqrt{2} C_1^* + 4) \rho (\frac{\lambda}{\kappa_\ell})^{1-q}. \qquad \text{(S3.17)}
\end{aligned}
$$

Since then, there exist constants C_6, C_7 and κ_ℓ, we have

$$
\| \widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}^* \|_F^2 \leq C_6 \rho \left(\frac{\lambda}{\kappa_\ell} \right)^{2-q}, \qquad \text{(S3.18)}
$$

$$
\| \widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}^* \|_* \leq C_7 \rho \left(\frac{\lambda}{\kappa_\ell} \right)^{1-q}. \qquad \text{(S3.19)}
$$

Put $\lambda = 2\nu\sqrt{dd_3/n}$ into (S3.18) and (S3.19), we have

$$\| \widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}^* \|_F^2 \leq C_6 \rho \left(\frac{\lambda}{\kappa_\ell} \right)^{2-q} = C_2 \rho \left(\frac{dd_3}{n} \right)^{1-q/2},$$
$$\| \widehat{\boldsymbol{\mathcal{M}}} - \boldsymbol{\mathcal{M}}^* \|_* \leq C_7 \rho \left(\frac{\lambda}{\kappa_\ell} \right)^{1-q} = C_3 \rho \left(\frac{dd_3}{n} \right)^{(1-q)/2},$$

where $C_2 = C_6(\frac{2\nu}{\kappa_\ell})^{2-q}, C_3 = C_7(\frac{2\nu}{\kappa_\ell})^{1-q}$ are constants.

Therefore, we finish the proof of Theorem 3.1. $\square$

S4. Proof of Lemma S3.1

Set $\eta_i^* = \langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle, \eta^* = \langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}} \rangle$. Since

$$\mathbb{E}[(b'(\eta^*) - \boldsymbol{Y})\boldsymbol{\mathcal{X}}] = \mathbb{E}[b'(\eta^*) - \boldsymbol{Y}]\mathbb{E}[\boldsymbol{\mathcal{X}}] = 0,$$

we have

$$\| \frac{1}{n} \sum_{i=1}^{n} (b'(\eta_i^*) - \boldsymbol{Y}_i)\boldsymbol{\mathcal{X}}_i \|_{op} = \| \frac{1}{n} \sum_{i=1}^{n} (b'(\eta_i^*) - \boldsymbol{Y}_i)\boldsymbol{\mathcal{X}}_i - \mathbb{E}[(b'(\eta^*) - \boldsymbol{Y})\boldsymbol{\mathcal{X}}] \|_{op} . \tag{S4.20}$$

Consider the following set $\mathcal{S}^{n-1} = \{\boldsymbol{u} \in \mathbb{R}^n : \| \boldsymbol{u} \|_2 = 1\}$. $\mathcal{N}^{n-1}$ denotes the 1/4 covering on $\mathcal{S}^{n-1}$, and $\Phi(\boldsymbol{\mathcal{A}}) = \sup_{\boldsymbol{u} \in \mathcal{N}^{dd_3-1}, \boldsymbol{v} \in \mathcal{N}^{dd_3-1}} \boldsymbol{u}^T bcirc(\boldsymbol{\mathcal{A}})\boldsymbol{v}$ for any $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d \times d \times d_3}$.

Since $\mathcal{N}^{dd_3-1}$ is the 1/4 covering on $\mathcal{S}^{dd_3-1}$, given $\boldsymbol{u} \in \mathcal{S}^{dd_3-1}, \boldsymbol{v} \in \mathcal{S}^{dd_3-1}$, there exist

$$\boldsymbol{u}_0 \in \mathcal{N}^{dd_3-1}, \boldsymbol{v}_0 \in \mathcal{N}^{dd_3-1},$$

which satisfy $\| \boldsymbol{u} - \boldsymbol{u}_0 \| \leq \frac{1}{4}, \| \boldsymbol{v} - \boldsymbol{v}_0 \| \leq \frac{1}{4}$. So that, we have

$$\begin{aligned}
&\boldsymbol{u}^T bcirc(\boldsymbol{\mathcal{A}})\boldsymbol{v} \\
=&\boldsymbol{u}_0^T bcirc(\boldsymbol{\mathcal{A}})\boldsymbol{v}_0 + \boldsymbol{u}_0^T bcirc(\boldsymbol{\mathcal{A}})(\boldsymbol{v} - \boldsymbol{v}_0) \\
&+ (\boldsymbol{u} - \boldsymbol{u}_0)^T bcirc(\boldsymbol{\mathcal{A}})\boldsymbol{v}_0 + (\boldsymbol{u} - \boldsymbol{u}_0)^T bcirc(\boldsymbol{\mathcal{A}})(\boldsymbol{v} - \boldsymbol{v}_0) \\
\leq&\Phi(\boldsymbol{\mathcal{A}}) + \frac{1}{4} \| bcirc(\boldsymbol{\mathcal{A}}) \|_{op} + \frac{1}{4} \| bcirc(\boldsymbol{\mathcal{A}}) \|_{op} + \frac{1}{16} \| bcirc(\boldsymbol{\mathcal{A}}) \|_{op} \\
=&\Phi(\boldsymbol{\mathcal{A}}) + \frac{9}{16} \| \boldsymbol{\mathcal{A}} \|_{op},
\end{aligned} \tag{S4.21}$$

where

$$\begin{aligned}
\| \boldsymbol{\mathcal{A}} \|_{op} =& \| bcirc(\boldsymbol{\mathcal{A}}) \|_{op} \\
=& \sup_{\boldsymbol{u} \in \mathcal{N}^{dd_3-1}, \boldsymbol{v} \in \mathcal{N}^{dd_3-1}} \boldsymbol{u}^T bcirc(\boldsymbol{\mathcal{A}})\boldsymbol{v} \\
\leq& \Phi(\boldsymbol{\mathcal{A}}) + \frac{9}{16} \| \boldsymbol{\mathcal{A}} \|_{op} .
\end{aligned} \tag{S4.22}$$

Hence, we conclude that

$$\| \boldsymbol{\mathcal{A}} \|_{op} \leq \frac{16}{7}\Phi(\boldsymbol{\mathcal{A}}). \tag{S4.23}$$

Let $z_i = b'(\eta_i^*) - \boldsymbol{Y}_i$ and $a_i = \eta_i^*$. Because

$$\begin{aligned}
\mathbb{E}\exp(tz_i|\boldsymbol{\mathcal{X}}_i) &= \int_y c(y)\exp(\frac{a_i y - b(a_i)}{\phi})\exp(t(y - b'(a_i)))dy \\
&= \int_y c(y)\exp\frac{1}{\phi}[(a_i + \phi t)y - b(a_i + \phi t) \\
&\quad + b(a_i + \phi t) - b(a_i) - \phi t b'(a_i)]dy \\
&= \exp(\frac{b(a_i + \phi t) - b(a_i) - \phi t b'(a_i)}{\phi}) \\
&\leq \exp(\frac{\phi t b'(a_i + \phi t) - \phi t b'(a_i)}{\phi}) \\
&= \exp(\frac{\phi^2 t^2 b''(a_*)}{\phi}) \\
&\leq \exp(\phi M t^2),
\end{aligned}$$

where $a_* \in (a_i, a_i + \phi t)$ is a certain point, it yields

$$\| z_i \|_{\psi_2} \leq C_0\sqrt{\phi M}. \tag{S4.24}$$

where C_0 is a positive constant. Thus, for any given $\boldsymbol{u} \in \mathcal{N}^{dd_3-1}, \boldsymbol{v} \in \mathcal{N}^{dd_3-1}$, let $Z_i = \boldsymbol{u}^T bcirc(\boldsymbol{\mathcal{X}}_i)\boldsymbol{v}$, $Z = \boldsymbol{u}^T bcirc(\boldsymbol{\mathcal{X}})\boldsymbol{v}$, we have $|Z_i| \leq \|$

$bcirc(\boldsymbol{\mathcal{X}}_i)\ \|_{op}= \sup_{\|\boldsymbol{a}\|\leq 1, \boldsymbol{a}\in\mathbb{R}^{dd_3}}\ \|\ bcirc(\boldsymbol{\mathcal{X}}_i)\boldsymbol{a}\ \|_2$. So that $\|\ Z_i\ \|_{\psi_2}\leq\|\ vec(bcirc(\boldsymbol{\mathcal{X}}_i))\ \|_{\psi_2}\leq \kappa_0$, and $\|\ (b'(\eta_i^*)-\boldsymbol{Y}_i)Z_i\ \|_{\psi_1}\leq\|\ b'(\eta_i^*)-\boldsymbol{Y}_i\ \|_{\psi_2}\|\ Z_i\ \|_{\psi_2}\leq C_0\sqrt{\phi M}\kappa_0$. By an application of Proposition 5.16 (Bermstein-type inequality) in Vershynin [2010], for sufficiently small t, it follows the following inequality

$$\mathbb{P}\bigg(\bigg|\frac{1}{n}\sum_{i=1}^{n}(b'(\eta_i^*)-\boldsymbol{Y}_i)Z_i-\mathbb{E}[(b'(\eta^*)-\boldsymbol{Y})Z]\bigg|>t\bigg)\leq 2\exp\bigg(-\frac{c_1nt^2}{\phi M\kappa_0^2}\bigg), \tag{S4.25}$$

where c_1 is a positive constant.

Combining the union bound over all points on $\mathcal{N}^{dd_3-1}\times\mathcal{N}^{dd_3-1}$ in (S4.25) and the results in (S4.23), we get

$$\begin{aligned}
&\mathbb{P}\bigg(\max_{\substack{\boldsymbol{u}\in\mathcal{N}^{dd_3-1}\\ \boldsymbol{v}\in\mathcal{N}^{dd_3-1}}}\bigg|\frac{1}{n}\sum_{i=1}^{n}(b'(\eta_i^*)-\boldsymbol{Y}_i)Z_i-\mathbb{E}[(b'(\eta^*)-\boldsymbol{Y})Z]\bigg|>t\bigg)\\
&\quad\leq\sum_{\boldsymbol{u}\in\mathcal{N}^{dd_3-1},\boldsymbol{v}\in\mathcal{N}^{dd_3-1}}2\exp\bigg(-\frac{c_1nt^2}{\phi M\kappa_0^2}\bigg)\\
&\quad\leq 2|\mathcal{N}^{dd_3-1}|^2\exp\bigg(-\frac{c_1nt^2}{\phi M\kappa_0^2}\bigg),
\end{aligned}\tag{S4.26}$$

where $|\mathcal{N}^{dd_3-1}|$ is the cardinality of $\mathcal{N}^{dd_3-1}$. According to Corollary 4.2.13 in Vershynin [2018], for our $\frac{1}{4}$-net $\mathcal{N}^{dd_3-1}$, we have $|\mathcal{N}^{dd_3-1}|\leq 9^{dd_3}$. Then the above inequality in (S4.26) can be deflated as

$$\mathbb{P}\bigg(\max_{\substack{\boldsymbol{u}\in\mathcal{N}^{dd_3-1}\\ \boldsymbol{v}\in\mathcal{N}^{dd_3-1}}} \bigg| \frac{1}{n}\sum_{i=1}^{n}(b'(\eta_i^*)-\boldsymbol{Y}_i)Z_i - \mathbb{E}[(b'(\eta^*)-\boldsymbol{Y})Z] \bigg| > t \bigg)$$
$$\le 2\cdot 9^{2dd_3}\exp\bigg(-\frac{c_1 n t^2}{\phi M \kappa_0^2}\bigg)$$
$$= 2\exp\bigg(2dd_3\log 9 - \frac{c_1 n t^2}{\phi M \kappa_0^2}\bigg). \tag{S4.27}$$

Furthermore, combining the above inequalities (S4.23) and (S4.27), we have

$$\mathbb{P}\bigg(\| \frac{1}{n}\sum_{i=1}^{n}(b'(\eta_i^*)-\boldsymbol{Y}_i)\boldsymbol{\mathcal{X}}_i \|_{op} > \frac{16}{7}t \bigg)$$
$$= \mathbb{P}\bigg(\| \frac{1}{n}\sum_{i=1}^{n}(b'(\eta_i^*)-\boldsymbol{Y}_i)bcirc(\boldsymbol{\mathcal{X}}_i) \|_{op} > \frac{16}{7}t \bigg)$$
$$= \mathbb{P}\bigg(\| \frac{1}{n}\sum_{i=1}^{n}(b'(\eta_i^*)-\boldsymbol{Y}_i)Z_i - \mathbb{E}[(b'(\eta^*)-\boldsymbol{Y})Z] \|_{op} > \frac{16}{7}t \bigg)$$
$$\le \mathbb{P}\bigg(\max_{\boldsymbol{u},\boldsymbol{v}} \bigg| \frac{1}{n}\sum_{i=1}^{n}(b'(\eta_i^*)-\boldsymbol{Y}_i)Z_i - \mathbb{E}[(b'(\eta^*)-\boldsymbol{Y})Z] \bigg| > t \bigg)$$
$$\le 2\exp\bigg(2dd_3\log 9 - \frac{c_1 n t^2}{\phi M \kappa_0^2}\bigg), \tag{S4.28}$$

where $\boldsymbol{u},\boldsymbol{v}\in\mathcal{N}^{dd_3-1}$.

Choosing $t \asymp \sqrt{\frac{dd_3}{n}}$ and selecting a constant $\gamma > 0$ such that $\frac{dd_3}{n} < \gamma$, it is easy to obtain the following inequality

$$\mathbb{P}\left(\| \frac{1}{n}\sum_{i=1}^{n}(b'(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle) - \boldsymbol{Y}_i)\cdot \boldsymbol{\mathcal{X}}_i \|_{op} > \nu\sqrt{\frac{dd_3}{n}}\right) \leq C\exp(-cdd_3). \tag{S4.29}$$

We prove the theorem completely. □

S5. Proof of Lemma S3.2

We will complete the proof divided into three parts as:

1. For all $\boldsymbol{\mathcal{D}} \in \mathbb{R}^{d\times d\times d_3}$ with probability greater than $1-\exp(-c_1 dd_3)$:

$$vec(\boldsymbol{\mathcal{D}})^T \cdot \hat{\boldsymbol{H}}(\boldsymbol{\mathcal{M}}^*)\cdot vec(\boldsymbol{\mathcal{D}}) \geq \kappa \cdot \| \boldsymbol{\mathcal{D}} \|_F^2 - C_0 d_3^2\sqrt{\frac{dd_3}{n}} \| \boldsymbol{\mathcal{D}} \|_*^2 . \tag{S5.30}$$

2. RSC at $L_n(\boldsymbol{\mathcal{M}}^*)$ over $\mathcal{C}(\mathcal{W}, \overline{\mathcal{W}}^{\perp}, \boldsymbol{\mathcal{M}}^*)$.

3. LRSC of $L_n(\boldsymbol{\mathcal{M}})$ around $\boldsymbol{\mathcal{M}}^*$.

Proof of Part 1:

As $\boldsymbol{\mathcal{D}} \to 0$, we have

$$
\begin{aligned}
L(\boldsymbol{\mathcal{M}}^* + \boldsymbol{\mathcal{D}}) - L(\boldsymbol{\mathcal{M}}^*) - \langle \nabla L(\boldsymbol{\mathcal{M}}^*), \boldsymbol{\mathcal{D}} \rangle =& \frac{1}{n} \sum_{i=1}^{n} b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle) \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{D}} \rangle^2 \\
=& vec(\boldsymbol{\mathcal{D}})^T \cdot \hat{\boldsymbol{H}}(\boldsymbol{\mathcal{M}}^*) \cdot vec(\boldsymbol{\mathcal{D}}),
\end{aligned}
$$

where $\boldsymbol{\mathcal{D}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^*$ denotes the t-SVD of $\boldsymbol{\mathcal{D}}$. We can easily obtain that $\| vec(\boldsymbol{\mathcal{S}}) \|_2 = \| \boldsymbol{\mathcal{D}} \|_F$. So that

$$
\begin{aligned}
& vec(\boldsymbol{\mathcal{D}})^T \cdot \hat{\boldsymbol{H}}(\boldsymbol{\mathcal{M}}^*) \cdot vec(\boldsymbol{\mathcal{D}}) \\
=& \frac{1}{n} \sum_{i=1}^{n} b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle) \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{D}} \rangle^2 \\
=& \frac{1}{n} \sum_{i=1}^{n} \left[\sqrt{b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle)} \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^* \rangle \right]^2 \\
=& \frac{1}{n} \sum_{i=1}^{n} \left[\sqrt{b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle)} \langle \boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{X}}_i * \boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{S}} \rangle \right]^2 \\
=& vec(\boldsymbol{\mathcal{S}})^T \cdot \widehat{\Sigma}_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} \cdot vec(\boldsymbol{\mathcal{S}}) \\
=& vec(\boldsymbol{\mathcal{S}})^T \cdot \Sigma_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} \cdot vec(\boldsymbol{\mathcal{S}}) + vec(\boldsymbol{\mathcal{S}})^T \cdot (\widehat{\Sigma}_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} - \Sigma_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}}) \cdot vec(\boldsymbol{\mathcal{S}}),
\end{aligned}
\tag{S5.31}
$$

where $\tilde{\boldsymbol{\mathcal{X}}}_i = \sqrt{b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle)} \cdot \boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{X}}_i * \boldsymbol{\mathcal{V}}$, $\widehat{\Sigma}_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} = \frac{1}{n} \sum_{i=1}^{n} vec(\tilde{\boldsymbol{\mathcal{X}}}_i) \cdot vec(\tilde{\boldsymbol{\mathcal{X}}}_i)^T$, and $\Sigma_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} = \mathbb{E}\widehat{\Sigma}_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}}$. In addition, we have

$$
\begin{aligned}
\lambda_{\min}(\Sigma_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}) = & \inf_{\substack{\boldsymbol{\mathcal{Z}}_1,\boldsymbol{\mathcal{Z}}_2 \in \mathbb{R}^{d\times d\times d_3} \\ \|\boldsymbol{\mathcal{Z}}_1\|_F = \|\boldsymbol{\mathcal{Z}}_2\|_F = 1}} vec(\boldsymbol{\mathcal{Z}}_1)^T \Sigma_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} vec(\boldsymbol{\mathcal{Z}}_2) \\
= & \inf \mathbb{E}\left[b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle)\langle \boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{X}}_i * \boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{Z}}_1\rangle \langle \boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{X}}_i * \boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{Z}}_2\rangle\right] \\
= & \inf \mathbb{E}[b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle)\langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{Z}}_1 * \boldsymbol{\mathcal{V}}^*\rangle \langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{Z}}_2 * \boldsymbol{\mathcal{V}}^*\rangle] \\
= & \inf vec(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{Z}}_1 * \boldsymbol{\mathcal{V}}^*)^T \cdot \hat{\boldsymbol{H}}(\boldsymbol{\mathcal{M}}^*) \cdot vec(\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{Z}}_2 * \boldsymbol{\mathcal{V}}^*) \\
= & \lambda_{\min}(\hat{\boldsymbol{H}}(\boldsymbol{\mathcal{M}}^*)) \geq \kappa. \qquad \text{(S5.32)}
\end{aligned}
$$

Therefore, according to Lemma S2.6, we have

$$
\begin{aligned}
& vec(\boldsymbol{\mathcal{D}})^T \cdot \hat{\boldsymbol{H}}(\boldsymbol{\mathcal{M}}^*) \cdot vec(\boldsymbol{\mathcal{D}}) \\
\geq & \kappa \parallel \boldsymbol{\mathcal{D}} \parallel_F^2 - |vec(\boldsymbol{\mathcal{S}})^T \cdot (\widehat{\Sigma}_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} - \Sigma_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}}) \cdot vec(\boldsymbol{\mathcal{S}})| \\
\geq & \kappa \parallel \boldsymbol{\mathcal{D}} \parallel_F^2 - \parallel \widehat{\Sigma}_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} - \Sigma_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} \parallel_\infty \parallel \boldsymbol{\mathcal{S}} \parallel_1^2 \\
\geq & \kappa \parallel \boldsymbol{\mathcal{D}} \parallel_F^2 - c_t d_3^2 \parallel \widehat{\Sigma}_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} - \Sigma_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} \parallel_\infty \parallel \boldsymbol{\mathcal{D}} \parallel_*^2 . \qquad \text{(S5.33)}
\end{aligned}
$$

Let $\mathcal{S}^{d\times 1\times d_3} = \{\boldsymbol{\mathcal{U}} \in \mathbb{R}^{d\times 1\times d_3} : \parallel \boldsymbol{\mathcal{U}} \parallel_F = 1\}, \mathcal{N}^{d\times 1\times d_3}$ be the 1/8 covering on

$\mathcal{S}^{d\times 1\times d_3}$. For any $\boldsymbol{A} \in \mathbb{R}^{d^2 d_3 \times d^2 d_3}$, define

$$\Phi(\boldsymbol{A}) := \sup_{\substack{\boldsymbol{\mathcal{U}}_1, \boldsymbol{\mathcal{U}}_2 \in \mathcal{S}^{d\times 1\times d_3} \\ \boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{V}}_2 \in \mathcal{S}^{d\times 1\times d_3}}} vec(\boldsymbol{\mathcal{U}}_1 * \boldsymbol{\mathcal{V}}_1^*)^T \cdot \boldsymbol{A} \cdot vec(\boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{V}}_2^*),$$

$$\Phi_{\mathcal{N}}(\boldsymbol{A}) := \sup_{\substack{\boldsymbol{\mathcal{U}}_1, \boldsymbol{\mathcal{U}}_2 \in \mathcal{N}^{d\times 1\times d_3} \\ \boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{V}}_2 \in \mathcal{N}^{d\times 1\times d_3}}} vec(\boldsymbol{\mathcal{U}}_1 * \boldsymbol{\mathcal{V}}_1^*)^T \cdot \boldsymbol{A} \cdot vec(\boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{V}}_2^*).$$

Applying the similar argument as that to prove the results in (S4.21) and (S4.23), we conclude that

$$vec(\boldsymbol{\mathcal{U}}_1 * \boldsymbol{\mathcal{V}}_1^*)^T \cdot \boldsymbol{A} \cdot vec(\boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{V}}_2^*) \leq \Phi_{\mathcal{N}}(\boldsymbol{A}) + \frac{9}{16}\Phi(\boldsymbol{A}). \tag{S5.34}$$

Thus, we have

$$\Phi(\boldsymbol{A}) \leq (16/7)\Phi_{\mathcal{N}}(\boldsymbol{A}) \tag{S5.35}$$

Lemma 5.14 in Vershynin [2010] yields that for any $\boldsymbol{\mathcal{U}}_1, \boldsymbol{\mathcal{U}}_2, \boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{V}}_2 \in \mathcal{S}^{d\times 1\times d_3}$, we have

$$
\begin{aligned}
&\| \langle \boldsymbol{\mathcal{U}}_1 * \boldsymbol{\mathcal{V}}_1^*, \tilde{\boldsymbol{\mathcal{X}}}_i \rangle \langle \boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{V}}_2^*, \tilde{\boldsymbol{\mathcal{X}}}_i \rangle \|_{\psi_1} \\
\leq & \frac{1}{2} (\| \langle \boldsymbol{\mathcal{U}}_1 * \boldsymbol{\mathcal{V}}_1^*, \tilde{\boldsymbol{\mathcal{X}}}_i \rangle^2 \|_{\psi_1} + \| \langle \boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{V}}_2^*, \tilde{\boldsymbol{\mathcal{X}}}_i \rangle^2 \|_{\psi_1}) \\
\leq & \| \langle \boldsymbol{\mathcal{U}}_1 * \boldsymbol{\mathcal{V}}_1^*, \sqrt{b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle)} \cdot \boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{X}}_i * \boldsymbol{\mathcal{V}} \rangle \|_{\psi_2}^2 \\
& + \| \langle \boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{V}}_2^*, \sqrt{b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle)} \cdot \boldsymbol{\mathcal{U}}^* * \boldsymbol{\mathcal{X}}_i * \boldsymbol{\mathcal{V}} \rangle \|_{\psi_2}^2 \\
\leq & 2M\kappa_0^2.
\end{aligned}
\tag{S5.36}
$$

Applying Bernstein Inequality in Vershynin [2010], we have

$$
\begin{aligned}
&\mathbb{P}\bigg(|vec(\boldsymbol{\mathcal{U}}_1 * \boldsymbol{\mathcal{V}}_1^*)^T (\widehat{\Sigma}_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} - \Sigma_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}}) vec(\boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{V}}_2^*)| > t \bigg) \\
&\quad \leq 2 \exp \bigg[- c \min(\frac{nt^2}{M^2\kappa_0^4}, \frac{nt}{M\kappa_0^2}) \bigg].
\end{aligned}
\tag{S5.37}
$$

At last, when we vectorized the tensors, the union bound over $(\boldsymbol{\mathcal{U}}_1, \boldsymbol{\mathcal{U}}_2, \boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{V}}_2) \in \mathcal{N}^{d\times 1\times d_3} \times \mathcal{N}^{d\times 1\times d_3} \times \mathcal{N}^{d\times 1\times d_3} \times \mathcal{N}^{d\times 1\times d_3}$ can be seen as the union bound over $(vec(\boldsymbol{\mathcal{U}}_1), vec(\boldsymbol{\mathcal{U}}_2), vec(\boldsymbol{\mathcal{V}}_1), vec(\boldsymbol{\mathcal{V}}_2)) \in \mathcal{N}^{dd_3-1} \times \mathcal{N}^{dd_3-1} \times \mathcal{N}^{dd_3-1} \times \mathcal{N}^{dd_3-1}$.

Hence, we have

$$
\begin{aligned}
&\mathbb{P}\bigg(|\sup vec(\boldsymbol{\mathcal{U}}_1 * \boldsymbol{\mathcal{V}}_1^*)^T(\widehat{\Sigma}_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} - \Sigma_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}})vec(\boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{V}}_2^*)| > t\bigg) \\
\leq \quad & 2\exp\bigg[4dd_3 \log 17 - c\min\big(\frac{nt^2}{M^2\kappa_0^4}, \frac{nt}{M\kappa_0^2}\big)\bigg], \qquad \text{(S5.38)}
\end{aligned}
$$

where $\boldsymbol{\mathcal{U}}_1, \boldsymbol{\mathcal{U}}_2 \in \mathcal{S}^{d\times 1\times d_3}, \boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{V}}_2 \in \mathcal{S}^{d\times 1\times d_3}$.

Choosing $t \asymp \sqrt{\frac{dd_3}{n}}$, we have that for some appropriate constants c_3, c_4 and C_1,

$$
\mathbb{P}\bigg(|\sup vec(\boldsymbol{\mathcal{U}}_1 * \boldsymbol{\mathcal{V}}_1^*)^T(\widehat{\Sigma}_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} - \Sigma_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}})vec(\boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{V}}_2^*)| > C_1\sqrt{\frac{dd_3}{n}}\bigg) \leq c_3\exp(-c_4 dd_3), \qquad \text{(S5.39)}
$$

where $\boldsymbol{\mathcal{U}}_1, \boldsymbol{\mathcal{U}}_2 \in \mathcal{S}^{d\times 1\times d_3}, \boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{V}}_2 \in \mathcal{S}^{d\times 1\times d_3}$.

Combining (S5.33) and (S5.39), we can conclude that

$$
\mathbb{P}\bigg(\parallel \widehat{\Sigma}_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} - \Sigma_{\tilde{\boldsymbol{\mathcal{X}}}\tilde{\boldsymbol{\mathcal{X}}}} \parallel_\infty > C_1\sqrt{\frac{dd_3}{n}}\bigg) \leq c_3\exp(-c_4 dd_3). \qquad \text{(S5.40)}
$$

We complete the proof of the first part.

Proof of Part 2:

Recall the definition of $\mathcal{C}(\mathcal{W}_r, \overline{\mathcal{W}}_r^{\perp}, \boldsymbol{\mathcal{M}}^*)$ as:

$$\mathcal{C}(\mathcal{W}_r, \overline{\mathcal{W}}_r^{\perp}, \boldsymbol{\mathcal{M}}^*) := \{\| \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}_r^{\perp}} \|_* \leq 3 \| \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}_r} \|_* + 4 \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*)\},$$

where $1 \leq r \leq d$. For any $\boldsymbol{\mathcal{D}} \in \mathcal{C}(\mathcal{W}_r, \overline{\mathcal{W}}_r^{\perp}, \boldsymbol{\mathcal{M}}^*)$, we have

$$\begin{aligned} \| \boldsymbol{\mathcal{D}} \|_* &\leq \| \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}_r} \|_* + \| \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}_r^{\perp}} \|_* \\ &\leq 4 \| \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}_r} \|_* + 4 \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*) \\ &\leq 4\sqrt{2r} \| \boldsymbol{\mathcal{D}}_{\overline{\mathcal{W}}_r} \|_F + 4 \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*) \\ &\leq 4\sqrt{2r} \| \boldsymbol{\mathcal{D}} \|_F + 4 \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*). \end{aligned} \tag{S5.41}$$

Set $\tilde{\kappa} = (1/8)\kappa$. Taking $\tau = \lambda/\tilde{\kappa}$ and $r = \#\{j \in \{1, 2, \ldots, d\} | \sigma_j(\boldsymbol{\mathcal{M}}^*) \geq \tau\}$, we get

$$\sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*) \leq \tau^{1-q} \sum_{j \geq r+1} \sigma_j(\boldsymbol{\mathcal{M}}^*)^q \leq \tau^{1-q}\rho = \lambda^{1-q}\tilde{\kappa}^{q-1}\rho. \tag{S5.42}$$

At the same time, $\rho \geq \sum_{j \leq r} \sigma_j(\boldsymbol{\mathcal{M}}^*)^q \geq r\tau^q$, so $r \leq \rho\tau^{-q} = \rho\tilde{\kappa}^q\lambda^{-q}$. Put these results into (S5.41), we have

$$\| \boldsymbol{\mathcal{D}} \|_* \leq 4\sqrt{2\rho}\lambda^{-q/2}\tilde{\kappa}^{q/2} \| \boldsymbol{\mathcal{D}} \|_F + 4\lambda^{1-q}\tilde{\kappa}^{q-1}\rho. \tag{S5.43}$$

Choosing $\lambda = 2\nu\sqrt{\frac{dd_3}{n}}$, we have $\rho d_3^2(dd_3/n)^{(1-q)/2} \leq C_6$ for some constants c_5 and C_6. Considering (S5.30) and (S5.43), we have

$$
\begin{aligned}
& vec(\boldsymbol{\mathcal{D}})^T \cdot \hat{\boldsymbol{H}}(\boldsymbol{\mathcal{M}}^*) \cdot vec(\boldsymbol{\mathcal{D}}) \\
\geq & \kappa \cdot \| \boldsymbol{\mathcal{D}} \|_F^2 - C_0 d_3^2 \sqrt{\frac{dd_3}{n}} \left[2(4\sqrt{2\rho}\lambda^{-q/2}\tilde{\kappa}^{q/2} \| \boldsymbol{\mathcal{D}} \|_F)^2 + 2(4\lambda^{1-q}\tilde{\kappa}^{q-1}\rho)^2 \right] \\
= & (\kappa - C_{10} d_3^2 \lambda^{1-q} \rho \tilde{\kappa}^q) \| \boldsymbol{\mathcal{D}} \|_F^2 - C_{20} \tilde{\kappa}^{2q-2} \lambda^{3-2q} \rho^2 \\
= & (\kappa - C_{11} d_3^2 \lambda^{1-q} \rho) \| \boldsymbol{\mathcal{D}} \|_F^2 - C_{21} \lambda^{3-2q} \rho^2 d_3^2 \\
\geq & (\kappa - C_{11} d_3^2 \lambda^{1-q} \rho) \| \boldsymbol{\mathcal{D}} \|_F^2 - C_{21} C_6 \lambda^{2-q} \rho d_3^2 \\
\geq & \tilde{\kappa} \| \boldsymbol{\mathcal{D}} \|_F^2 - c_5 d_3^2 \rho \lambda^{2-q},
\end{aligned}
\tag{S5.44}
$$

with high probability, where $\rho(dd_3/n)^{(1-q)/2} \leq C_6$ holds for $\kappa - C_{11} d_3^2 \lambda^{1-q} \rho \geq \tilde{\kappa}$. We complete the proof of the second part.

Before we present the proof of the third part, we establish the following lemma.

Lemma S5.1. *If* $\| \boldsymbol{\mathcal{M}}^* \|_F \geq \alpha\sqrt{dd_3}$ *and* $\{vec(\boldsymbol{\mathcal{X}}_i)\}_{i=1}^n$ *are sub-Gaussian,* $\lambda_{min}(\boldsymbol{h}(\boldsymbol{\mathcal{M}}^*)) \geq \kappa_1$, *where* κ_1 *is a positive constant, for a universal constant* $\tau > 0$.

Proof:

Because $\langle \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{X}}_i \rangle$ is a linear transformation of a sub-Gaussian vector, it is a sub-Gaussian variable, and its mean is 0 and its sub-Gaussian norm is bounded by $\kappa_0 \parallel \boldsymbol{\mathcal{M}} \parallel_F$. When $\parallel \boldsymbol{\mathcal{M}} \parallel_F \geq \alpha\sqrt{dd_3}$, taking a sufficiently small c_1, we have

$$\begin{aligned}\mathbb{P}(|\langle \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{X}}_i\rangle| > c_1\sqrt{dd_3}) \geq & \mathbb{P}(|\langle \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{X}}_i\rangle| > c_1 \parallel \boldsymbol{\mathcal{M}} \parallel_F /\alpha) \\ = & \mathbb{P}(|\langle \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{X}}_i\rangle| / \parallel \boldsymbol{\mathcal{M}} \parallel_F > c_1/\alpha) \\ \geq & \frac{p_0+1}{2}. \end{aligned} \tag{S5.45}$$

Additionally, we have

$$\begin{aligned}\parallel \boldsymbol{\mathcal{X}}_i \parallel_{op} = & \parallel bcirc(\boldsymbol{\mathcal{X}}_i) \parallel_{op} \\ = & \max_{\boldsymbol{u}\in\mathcal{S}^{dd_3-1}, \boldsymbol{v}\in\mathcal{S}^{dd_3-1}} |\langle \boldsymbol{u}, bcirc(\boldsymbol{\mathcal{X}}_i)\boldsymbol{v}\rangle| \\ = & \max_{\boldsymbol{u}\in\mathcal{S}^{dd_3-1}, \boldsymbol{v}\in\mathcal{S}^{dd_3-1}} |\langle \boldsymbol{u}\boldsymbol{v}^T, bcirc(\boldsymbol{\mathcal{X}}_i)\rangle|.\end{aligned}$$

Let $\mathcal{N}^{dd_3-1}$ denote a 1/4-net on $\mathcal{S}^{dd_3-1}$, and then we have

$$\max_{\substack{\boldsymbol{u}\in\mathcal{S}^{dd_3-1}\\ \boldsymbol{v}\in\mathcal{S}^{dd_3-1}}} |\langle \boldsymbol{u}\boldsymbol{v}^T, bcirc(\boldsymbol{\mathcal{X}}_i)\rangle| \leq \frac{16}{7} \max_{\substack{\boldsymbol{u}\in\mathcal{S}^{dd_3-1}\\ \boldsymbol{v}\in\mathcal{S}^{dd_3-1}}} |\langle \boldsymbol{u}\boldsymbol{v}^T, bcirc(\boldsymbol{\mathcal{X}}_i)\rangle|.$$

Since $\parallel vec(\boldsymbol{\mathcal{X}}_i) \parallel_{\psi_2} \leq \kappa_0$, we have$\parallel \langle \boldsymbol{u}\boldsymbol{v}^T, bcirc(\boldsymbol{\mathcal{X}}_i)\rangle \parallel_{\psi_2} \leq \kappa_0$ for any $\boldsymbol{u} \in \mathcal{N}^{dd_3-1}, \boldsymbol{v} \in \mathcal{N}^{dd_3-1}$. According to Definition 2.5.6 in Vershynin [2018], for some sufficiently small t and some positive constant C, we have

$$\mathbb{P}(|\langle \boldsymbol{u}\boldsymbol{v}^T, bcirc(\boldsymbol{\mathcal{X}}_i)\rangle| > t) \leq 2\exp\left(-\frac{Ct^2}{\kappa_0^2}\right).$$

Hence, we have

$$\mathbb{P}\left(\max_{\boldsymbol{u}\in\mathcal{S}^{dd_3-1},\boldsymbol{v}\in\mathcal{S}^{dd_3-1}} |\langle \boldsymbol{u}\boldsymbol{v}^T, bcirc(\boldsymbol{\mathcal{X}}_i)\rangle| > t\right) \leq 2\exp\left(2dd_3\log 9 - \frac{Ct^2}{\kappa_0^2}\right).$$

For some positive constant $c_2 > \sqrt{4\log 9\kappa_0^2/C}$, we choose $t = c_2\sqrt{dd_3}$ and obtain that

$$\mathbb{P}(\| \boldsymbol{\mathcal{X}}_i \|_{op} \leq c_2\sqrt{dd_3}) \geq (p_0+1)/2. \tag{S5.46}$$

Therefore, altogether (S5.45) and (S5.46), we have

$$\begin{aligned}
&\mathbb{P}(|\langle \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{X}}_i\rangle| > c_1/c_2 \| \boldsymbol{\mathcal{X}}_i \|_{op}) \\
\geq&\mathbb{P}(|\langle \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{X}}_i\rangle| > c_1\sqrt{dd_3}) + \mathbb{P}(\| \boldsymbol{\mathcal{X}}_i \|_{op} \leq c_2\sqrt{dd_3}) \\
\geq&(p_0+1)/2 + (p_0+1)/2 - 1 = p_0.
\end{aligned}$$

Because $vec(\boldsymbol{\mathcal{X}}_i)$ is sub-Gaussian, the fourth moment is bounded. For any $\boldsymbol{v} \in \mathbb{R}^{d^2 d_3}$, we have

$$
\begin{aligned}
& n\boldsymbol{v}^T \boldsymbol{h}(\boldsymbol{\mathcal{M}}^*)\boldsymbol{v} \\
=& \mathbb{E}\Big[\sum_{i=1}^{n} b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle)(vec(\boldsymbol{\mathcal{X}}_i)^T\boldsymbol{v})^2\Big] - \mathbb{E}\Big[\sum_{i=1}^{n} b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle)\cdot \mathbf{1}_{A_i^c}\cdot (vec(\boldsymbol{\mathcal{X}}_i)^T\boldsymbol{v})^2\Big] \\
\geq& n\boldsymbol{v}^T \hat{\boldsymbol{H}}(\boldsymbol{\mathcal{M}}^*)\boldsymbol{v} - \sqrt{\mathbb{E}\Big[\sum_{i=1}^{n} b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle)^2 (vec(\boldsymbol{\mathcal{X}}_i)^T\boldsymbol{v})^4\Big]}\cdot\sqrt{\mathbb{E}\sum_{i=1}^{n}\mathbf{1}_{A_i^c}} \\
\geq& n\kappa \parallel \boldsymbol{v} \parallel_2^2 - nMK\sqrt{1-p_0} \parallel \boldsymbol{v} \parallel_2^2, \qquad \text{(S5.47)}
\end{aligned}
$$

where M is an global upper bound of $b''(\cdot)$ and K is the largest eigenvalue of the fourth moment of $\boldsymbol{\mathcal{X}}_i$. Let $1-p_0$ to be small enough that $MK\sqrt{1-p_0} \leq \kappa/2$. Then we can easily prove that $\lambda_{\min}(\boldsymbol{h}(\boldsymbol{\mathcal{M}}^*)) > \kappa/2 > 0$, and thus $\boldsymbol{h}(\boldsymbol{\mathcal{M}}^*)$ is positive definite. $\square$

Now we turn to present the third part of the justification in Lemma S3.2.

Proof of Part 3:

Define $\hat{\boldsymbol{h}}(\boldsymbol{\mathcal{M}}) = \frac{1}{n}\sum_{i=1}^{n} b''(\langle \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{X}}_i\rangle)\cdot \mathbf{1}_{\{|\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i\rangle| > \tau \|\boldsymbol{\mathcal{X}}_i\|_{op}\}}\cdot vec(\boldsymbol{\mathcal{X}}_i)vec(\boldsymbol{\mathcal{X}}_i)^T$ and $\boldsymbol{h}(\boldsymbol{\mathcal{M}}) = \mathbb{E}(\hat{\boldsymbol{h}}(\boldsymbol{\mathcal{M}}))$ for constants τ and γ.

Based on Lemma S5.1, we choose an appropriate τ to make $\boldsymbol{h}(\boldsymbol{\mathcal{M}}^*)$ positive definite. Following the similar description as that in Part 1 and

Part 2, we can easily prove that

$$vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot \hat{\boldsymbol{h}}(\boldsymbol{\mathcal{M}}^*) \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}) \geq \tilde{\kappa}_1 \parallel \widehat{\boldsymbol{\mathcal{D}}} \parallel_F^2 - c_6 d_3^2 \rho \lambda^{2-q}, \tag{S5.48}$$

for some positive $\tilde{\kappa}_1$ with high probability.

On the other hand, we have that

$$\begin{aligned}
&|vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot \hat{\boldsymbol{h}}(\boldsymbol{\mathcal{M}}^*) \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}) - vec(\hat{\boldsymbol{\mathcal{D}}})^T \cdot \widehat{\boldsymbol{h}}(\boldsymbol{\mathcal{M}}) \cdot vec(\widehat{\boldsymbol{\mathcal{D}}})| \\
\leq & vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot \frac{1}{n} \sum_{i=1}^n |b''(\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle) - b''(\langle \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{X}}_i \rangle)| \\
& \quad \cdot \mathbf{1}_{\{|\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle| > \tau \| \boldsymbol{\mathcal{X}}_i \|_{op}\}} \cdot vec(\boldsymbol{\mathcal{X}}_i) vec(\boldsymbol{\mathcal{X}}_i)^T \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}) \\
= & vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot \frac{1}{n} \sum_{i=1}^n |b'''(\langle \tilde{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{X}}_i \rangle) \langle \boldsymbol{\mathcal{M}} - \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle| \\
& \quad \cdot \mathbf{1}_{\{|\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle| > \tau \| \boldsymbol{\mathcal{X}}_i \|_{op}\}} \cdot vec(\boldsymbol{\mathcal{X}}_i) vec(\boldsymbol{\mathcal{X}}_i)^T \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}),
\end{aligned} \tag{S5.49}$$

where $\tilde{\boldsymbol{\mathcal{M}}}$ is a middle point between $\boldsymbol{\mathcal{M}}^*$ and $\boldsymbol{\mathcal{M}}$. Hence, $\tilde{\boldsymbol{\mathcal{M}}}$, which is centered at $\boldsymbol{\mathcal{M}}^*$ with radius ℓ, is also in the nuclear ball.

If the indicator function equals to 1, we have $|\langle \tilde{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{X}}_i \rangle| \geq |\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle| - |\langle \boldsymbol{\mathcal{M}}^* - \tilde{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{X}}_i \rangle| \geq (\tau - \ell) \parallel \boldsymbol{\mathcal{X}}_i \parallel_{op}$. According to Condition (5) in Lemma

S3.2, if $|\langle \tilde{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{X}}_i \rangle| > 1$, we have

$$
\begin{aligned}
&|b'''(\langle \tilde{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{X}}_i \rangle)\langle \boldsymbol{\mathcal{M}} - \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle| \cdot \mathbf{1}_{\{|\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle| > \tau \|\boldsymbol{\mathcal{X}}_i\|_{op}\}} \\
\leq & \frac{\| \boldsymbol{\mathcal{X}}_i \|_{op} \| \boldsymbol{\mathcal{M}} - \boldsymbol{\mathcal{M}}^* \|_*}{|\langle \tilde{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{X}}_i \rangle|} \\
\leq & \frac{\| \boldsymbol{\mathcal{X}}_i \|_{op} \| \boldsymbol{\mathcal{M}} - \boldsymbol{\mathcal{M}}^* \|_*}{(\tau - \ell) \| \boldsymbol{\mathcal{X}}_i \|_{op}} \\
\leq & \frac{\ell}{\tau - \ell}.
\end{aligned}
\tag{S5.50}
$$

Meanwhile, $\| \boldsymbol{\mathcal{X}}_i \|_{op}$ is bounded by $1/(\tau - \ell)$ and $|b'''(\langle \tilde{\boldsymbol{\mathcal{M}}}, \boldsymbol{\mathcal{X}}_i \rangle)\langle \boldsymbol{\mathcal{M}} - \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle| \cdot \mathbf{1}_{\{|\langle \boldsymbol{\mathcal{M}}^*, \boldsymbol{\mathcal{X}}_i \rangle| > \tau \|\boldsymbol{\mathcal{X}}_i\|_{op}\}} \leq C \cdot \| \boldsymbol{\mathcal{M}} - \boldsymbol{\mathcal{M}}^* \|_* \| \boldsymbol{\mathcal{X}}_i \|_{op} \leq C \cdot \frac{\ell}{\tau - \ell}$, where C is the upper bound of $b'''(x)$ for $|x| \leq 1$. Thus, we have

$$
\begin{aligned}
&|vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot \hat{\boldsymbol{h}}(\boldsymbol{\mathcal{M}}^*) \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}) - vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot \hat{\boldsymbol{h}}(\boldsymbol{\mathcal{M}}) \cdot vec(\widehat{\boldsymbol{\mathcal{D}}})| \\
\leq & vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot \frac{C_1 \ell}{n(\tau - \ell)} \sum_{i=1}^{n} vec(\boldsymbol{\mathcal{X}}_i) vec(\boldsymbol{\mathcal{X}}_i)^T \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}),
\end{aligned}
\tag{S5.51}
$$

where $C_1 = \max(C, 1)$.

Let $\widehat{\Sigma}_{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}} = \frac{1}{n} \sum_{i=1}^n vec(\boldsymbol{\mathcal{X}}_i) \cdot vec(\boldsymbol{\mathcal{X}}_i)^T$ and $\Sigma_{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}} = \mathbb{E}\widehat{\Sigma}_{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}}$. Denote the upper bound of the eigenvalues of $\Sigma_{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}}$ as $K (K < \infty)$, and recall the proof process of (S5.39), we have

$$
\begin{aligned}
&vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot \frac{C_1 l}{n(\tau - l)} \sum_{i=1}^{n} vec(\boldsymbol{\mathcal{X}}_i) vec(\boldsymbol{\mathcal{X}}_i)^T \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}) \\
=&\frac{C_1 \ell}{\tau - \ell} \cdot vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot \widehat{\Sigma}_{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}} \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}) \\
=&\frac{C_1 \ell}{\tau - \ell} \cdot \left[vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot \Sigma_{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}} \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}) + vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot (\widehat{\Sigma}_{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}} - \Sigma_{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}}) \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}) \right] \\
\leq&\frac{C_1 \ell}{\tau - \ell} \cdot \left[K \parallel \widehat{\boldsymbol{\mathcal{D}}} \parallel_F^2 + | vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot (\widehat{\Sigma}_{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}} - \Sigma_{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}}) \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}) | \right] \\
\leq&\frac{C_1 \ell}{\tau - \ell} \cdot \left[K \parallel \widehat{\boldsymbol{\mathcal{D}}} \parallel_F^2 + | \sup_{\substack{\boldsymbol{u} \in \mathcal{R}^{d^2 d_3} \\ \|\boldsymbol{u}\|_2 = 1}} \boldsymbol{u}^T \cdot (\widehat{\Sigma}_{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}} - \Sigma_{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}}) \cdot \boldsymbol{u} | \parallel \widehat{\boldsymbol{\mathcal{D}}} \parallel_F^2 \right] \\
\leq&\frac{C_1 \ell}{\tau - \ell} \cdot \left[K \parallel \widehat{\boldsymbol{\mathcal{D}}} \parallel_F^2 + C_0 \sqrt{\frac{d d_3}{n}} \parallel \widehat{\boldsymbol{\mathcal{D}}} \parallel_F^2 \right] \\
\leq&\frac{2 K C_1 \ell}{\tau - \ell} \parallel \widehat{\boldsymbol{\mathcal{D}}} \parallel_F^2,
\end{aligned} \tag{S5.52}
$$

with high probability.

In summary, if the constant ℓ is sufficiently small and satisfies $2KC_1\ell/(\tau-\ell) < \tilde{\kappa}_1/2$, we have $vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot \widehat{\boldsymbol{h}}(\boldsymbol{\mathcal{M}}) \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}) \geq \tilde{\kappa}_2 \parallel \widehat{\boldsymbol{\mathcal{D}}} \parallel_F^2$, where $\tilde{\kappa}_2 = \tilde{\kappa}_1/2$. This claims that $\widehat{\boldsymbol{h}}(\boldsymbol{\mathcal{M}})$ is locally positive definite around $\boldsymbol{\mathcal{M}}^*$ with high probability. Consider that $\boldsymbol{H}(\cdot) \geq \boldsymbol{h}(\cdot)$, we can easily prove that $\hat{\boldsymbol{H}}(\boldsymbol{\mathcal{M}})$ is also locally positive definite around $\boldsymbol{\mathcal{M}}^*$. Therefore, there exists some constant $\ell > 0$, for any $\| \boldsymbol{\mathcal{M}} - \boldsymbol{\mathcal{M}}^* \|_* \leq \ell$, we have

$$vec(\widehat{\boldsymbol{\mathcal{D}}})^T \cdot \frac{1}{n} \sum_{i=1}^{n} b''(\langle \boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{X}}_i \rangle) vec(\boldsymbol{\mathcal{X}}_i) vec(\boldsymbol{\mathcal{X}}_i)^T \cdot vec(\widehat{\boldsymbol{\mathcal{D}}}) \geq \tilde{\kappa}_2 \| \widehat{\boldsymbol{\mathcal{D}}} \|_F^2 - c_6 d_3^2 \rho \lambda^{2-q}, \tag{S5.53}$$

for all $\widehat{\boldsymbol{\mathcal{D}}} \in \mathcal{C}(\mathcal{W}_r, \overline{\mathcal{W}}_r^{\perp}, \boldsymbol{\mathcal{M}}^*)$ with high probability.

The whole Lemma S3.2 is proved. □

References

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 46(2):814, 2018.

Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):925–938, 2020.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. 2018.